MISB MOTION IMAGERY STANDARDS BOARD	MISB EG 0802.1
Engineering Guideline	**30 Sep 2010**
H.264 / AVC Coding and Multiplexing	

1 Scope

This H.264/AVC (ITU-T Rec. H.264 | ISO/IEC 14496-10) Coding and Multiplexing Engineering Guide provides recommendations and guidelines for the creation of H.264/AVC Motion Imagery that maintains the interoperability and quality for motion imagery within the US Department of Defense / Intelligence Community / National System for Geospatial-Intelligence (DoD/IC/NSG) community.

The document covers the creation of H.264/AVC Motion Imagery for:

- operator-in-the-loop operations
- video exploitation and dissemination
- low bandwidth users

The scope of this document includes the structure of the video encoding, information included with the video, and properties of the transport layer used to carry the video.

2 References

[1] ITU-T Rec. H.264 (03/09), Advanced video coding for generic audiovisual service / ISO/IEC 14496-10:2009 *Information Technology - Coding of audio-visual objects Part 10: Advanced Video Coding*

[2] ISO/IEC 13818-4:2004. Information technology – *Generic coding of moving pictures and associated audio information: Systems -- Part 4: Conformance testing*

[3] ITU-T Rec.H.222 (05/06) | ISO/IEC 13818-1:2007. *Information technology – Generic coding of moving pictures and associated audio information: Systems*

[4] ITU-T Rec.H.222 (05/06) | ISO/IEC 13818-1:2007 Amendment 3: *Transport of AVC video data over ITU-T Rec. H.222.0 | ISO/IEC 13818-1 streams*

[5] ETSI TS 101 154 V1.8.1 (2005-06): Digital Video Broadcasting (DVB); *Implementation guidelines for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream*

[6] ETSI TS 102 005 V1.3.1 (draft): Digital Video Broadcasting (DVB); *Specification for the use of Video and Audio Coding in DVB services delivered directly over IP protocols*

[7] IETF RFC 3984: *RTP payload for transport of H.264*, Feb 2005

[8] SCTE 128 2008: *AVC Video Systems and Transport Constraints for Cable Television*

[9] MISB RP 0804.2, *Real-Time Protocol for Full Motion Video*, Jun 2010

[10] MISB STANDARD 0604.1, *Time Stamping Compressed Motion Imagery*, Sep 2009

3 Acronyms

ASO	Arbitrary Slice Ordering
AU	Access Unit
AVC	Advanced Video Coding (H264)
CABAC	Context-Adaptive Binary Arithmetic Coding
CAVLC	Context-Adaptive Variable Length Coding
COTS	Commercial Off-The-Shelf
CPB	Coded Picture Buffer
DCT	Discrete Cosine Transform
DPB	Decoded Picture Buffer
DTS	Decoding Time Stamp
DVB	Digital Video Broadcast
FMO	Flexible Macroblock Ordering
GOP	Group of Pictures
HRD	Hypothetical Reference Decoder
IDR	Instantaneous Decoding Refresh
JVT	Joint Video Team (ITU-T and ISO/IEC)
MP4	MPEG4
NAL	Network Abstraction Layer
PAT	Program Association Table
PCR	Program Clock Reference
PES	Packetized Elementary Stream
PMT	Program Map Table
POC	Picture Order Count
PPS	Picture Parameter Set
PTS	Presentation Time Stamp
PVR	Personal Video Recorder
RAP	Random Access Point
RTP	Real Time Protocol
SEI	Supplemental Enhancement Information
SPS	Sequence Parameter Set
VCL	Video Coding Layer
VUI	Video Usability Information

4 Introduction

The H.264/AVC video coding standard [1] is capable of delivering significant compression increases over the earlier MPEG-2 video standard; for a given bandwidth H.264/AVC is capable of delivering double the resolution, for a given resolution it may require only half the bandwidth. There are many advantages for migrating systems to use H.264/AVC and over time it will replace MPEG-2 in many domains. It already forms the basis of the next generation of DVD, digital satellite broadcast, and Digital Video Broadcast (DVB) standards.

Video is used for many different purposes within the US Department of Defense / Intelligence Community / National System for Geospatial-Intelligence (DoD/IC/NSG) community. Because different application areas have quite different requirements, it is difficult to define a single set of H.264/AVC encoding and multiplexing guidelines which will meet the requirements for all users across the DoD/IC/NSG.

The uses of video within the DoD/IC/NSG community can be broadly categorized as follows:

- Video for operator-in-the-loop operations (low latency)
- Video for exploitation and dissemination
- Video for low bandwidth users

For operator-in-the-loop operations, the most important video requirement is to be able to view the live video stream with a very low end-to-end latency and consistent video quality. Since the operator is trying to react in real time to imagery that is being viewed, it is critical that latency introduced by the motion imagery and communications systems is minimized. Studies have shown that the *glass-to-glass* latency needs to be less than 200 milliseconds. Typical operator activities include:

- Manually tracking a moving vehicle down a highway
- Driving an unmanned vehicle down a road
- Manipulating an Explosive Ordinance Disposal robotic arm

For video exploitation[1] and dissemination, the most important requirements are video quality and the ability to support random access and PVR play modes such as fast random access, frame stepping, fast forward, fast reverse and looping. Typical exploitation and dissemination activities include:

- Searching through and analyzing long video sequences
- Editing video clips from a video sequence
- Creating video files for distribution

[1] Phase 2 and Phase 3 video exploitation

For low bandwidth users, receiving real time video over the available bandwidth is the most important requirement. Typical low bandwidth user activities include:

- Target area observation
- Battle damage assessment

Due to the nature of the requirements of each application area, video encoded for one purpose may not always be suitable or even compatible with another. To help system designers and vendors create H.264/AVC video that best meets the needs of a specific purpose and to maximize compatibility with other users, this document presents a number of specific encoding and multiplexing guidelines.

First, a common profile is defined which provides a minimal set of encoding guidelines suitable for all three application areas. While not optimal for any one area, these guidelines provide a reasonable compromise and maximize compatibility across all types of systems. This is particularly important when video is used for sensor control and downstream exploitation and low-bandwidth dissemination.

For each of the three application areas, enhancements and changes from the common profile are described to create video that is further optimized for that area.

The relationship between the encoding parameters for each area can be shown using the Venn diagram in Figure 1 below. Each circle represents the parameters compatible for an application area. The intersection defines the common profile presented in this document.

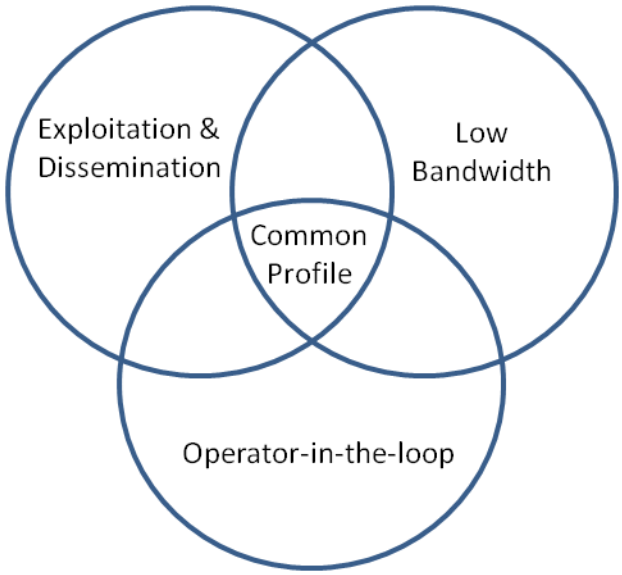

Figure 1 – Application areas typically require different encoding parameters

The conflict of requirements between areas can be explained using the triangle[2] of Quality, Bit rate and Latency as shown in Figure 2 below.

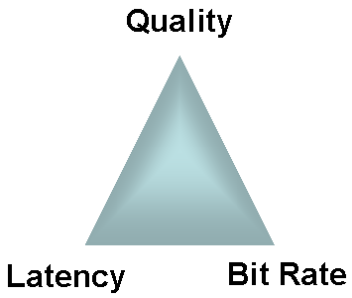

Figure 2 – Application requirements require tradeoffs in Quality, Latency, and Bit Rate

For Operator-in-the-loop operations, the latency of the video is one of the most important characteristics. Achieving low latency requires using encoding structures which are not the most efficient (e.g. no B frames and relatively fixed frame sizes), which reduces quality and increases the required bit rate.

For Video Exploitation and Dissemination, quality of the video is one of the most important characteristics. To get the highest quality for a given bit rate requires using B frames and a greater variation between frames sizes, both of which increase stream latency. Higher bit rates improve quality.

For the low bandwidth user, the bit rate of the video is the most important characteristic. To get the lowest bit rate requires using the most efficient encoding (which increases latency) and reduced quality.

By presenting clear engineering guidelines this document hopes to guide vendors in developing technologies and systems which are compatible with current and future systems used by the US Department of Defense / Intelligence Community / National System for Geospatial-Intelligence (DoD/IC/NSG) community.

The guidelines for the carriage of H.264/AVC in both MPEG-2 Transport Streams (Xon2) and the Real Time Protocol (RTP) are treated in this document since they are closely intertwined [9].

5 Coding Guidelines

The minimum requirements a video must satisfy to be compliant with the H.264/AVC standard only ensures there is enough information for every picture to be decodable in the right order. There are no specific requirements that mandate that information is carried to control the presentation or display of these pictures. The guidelines in the following sections are designed to maximize the usefulness of the encoded H.264/AVC video and increase its interoperability with

[2] The conflict in encoding parameters goes beyond a simple triangle representing three opposing areas, but a more detailed discussion is beyond the scope of this document

other systems. Many of the guidelines are derived from industry best practice rather than prescribed by the published standards.

5.1 Common H.264/AVC profile

To create a H.264/AVC stream that is sufficiently low latency to meet operator in the loop requirements, has enough information to support random access and trick mode playback for exploitation, and can be delivered to bandwidth challenged users, requires making many compromises. The set of encoding guidelines below represents a reasonable set of compromises and will produce a stream that has the greatest amount of interoperability amongst systems.

Table 1 - Recommendations for Common Profile

Guidelines	***Reference Section***
SPS and PPS occur before every IDR picture or the I picture at a recovery point	6.1
Random Access Points occur at least once a second	6.2
PTS values are present for every frame (Note: this is a systems, not a coding parameter)	6.3
Fixed frame rate (piece-wise fixed)	6.5
Fixed frame size (piece-wise fixed)	6.6
Picture Order Count (POC) equal to field number	6.8
IP coding structure (no B frames)	6.7
No long-term reference pictures	6.9
Use “picture not coded” when frames are dropped	6.11
Includes VUI parameters: *time_scale, num_units_in_ticks, fixed_frame_rate, num_reorder_frames, max_dec_frame_buffering*	6.12.1
Picture timing SEI message (STD 0604 [10])	6.13
Main or Constrained[3] Baseline Profile	6.10

5.2 Operator-in-the-loop H.264/AVC profile

The most important requirement for sensor or vehicle operators is latency. Since low latency encoding is less efficient than normal encoding modes, to maintain image quality over a given bandwidth requires careful attention to the contents of the stream.

The H.264/AVC encoding guidelines for operator-in-the-loop operations take the common guidelines and make the following additions and changes:

[3] Constrained Baseline Profile provides tools that are common between Baseline and Main profile. It excludes any of the error resilience tools.

Table 2 - Recommendations for Operator in the Loop

Guidelines	*Reference Section*
Frame rate (piece-wise fixed or variable)	6.5, 6.6
No B frames (I & P only)	6.7
Use RAP > 1 second or Infinite GOPs (P frame only[4])	6.2, 6.7
Baseline, Main or High Profile	6.7

In addition to supporting piecewise continuous fixed frame rate, low latency encoders may also support non-fixed or variable frame rate. Non-fixed frame rate allows frames to be encoded and transmitted as fast as they can. Note: this mode is incompatible with current exploitation systems.

5.3 Exploitation and Dissemination H.264/AVC profile

The three main requirements for motion imagery for exploitation and dissemination are compatibility, video quality, and the ability to support 'trick mode' playback which includes random access, fast forward and fast reverse.

The H.264/AVC encoding guidelines for exploitation and dissemination take the common guidelines and made the following additions and changes:

Table 3 - Recommendations for Exploitation and Dissemination

Guidelines	*Reference Section*
Use B frames (IBP, IBBP or IBBBP[5])	6.7
Include HRD Parameters: *cpb_size_value_minus1*	6.12.2
Include buffering period SEI message	6.12.3
Main or High Profile	6.10

The use of bi-directionally predicted frames (B frames, which can be either reference or non-reference/disposable) improves the quality of the video for a given bit rate. Disposable B frames also help support variable speed playback as they can be easily dropped during decoding. Because B frames introduce a frame re-ordering delay and increase the buffering requirements for constant bit rate streams, they are not suited for low-latency streams.

Reducing the distance between Random Access Points improves random access navigation, frame-accurate random access, fast forward and fast reverse playback.

[4] Currently incompatible with exploitation systems
[5] IBBBP refers to Hierarchical B frame encoding

5.4 Low Bandwidth Users H.264/AVC profile

Low bandwidth users over reliable or un-reliable communication links are defined as users who have insufficient useable bandwidth to receive the video and metadata at full frame rate and full resolution. To satisfy these requirements low bandwidth users have to trade off frame size, frame rate and the metadata frequency to fit the constraints of the available bandwidth.

The H.264/AVC encoding guidelines for low bandwidth users take the common guidelines and make the following additions and changes:

Table 4 - Recommendations for Low Bandwidth Users

Guidelines	*Reference Section*
Use B frames (IBP, IBBP or IBBBP[6])	6.7
RAP > 1sec or Infinite GOPs	6.2, 6.7
Error resilience tools	N/A
Baseline, Main or High Profile	6.7

Increasing the time between random access points using I or IDR frames can improve the quality of streams over lower bit rates, at the expense of longer times for decoders to open up and begin playing the stream, and longer times for error propagation.

Baseline profile provides a number of tools to provide error resilience which may be useful for low bandwidth users over lossy networks.

For low power portable devices, it may be necessary to use baseline or simple baseline profile, where only CAVLC entropy coding is supported.

6 Coding Details

6.1 SPS and PPS occur before an IDR picture or the I picture at a recovery point

The Video Coding Layer (VCL), which consists of a hybrid of temporal and spatial prediction in conjunction with transform coding, is specified to efficiently represent the content of the video data. The Network Abstraction Layer (NAL) is specified to format that data and provide header information in a manner appropriate for conveyance by the transport layers or storage media. All data are contained in NAL units, each of which contains an integer number of bytes. The NAL facilitates the ability to map H.264/AVC VCL data to transport layers such as:

- RTP/IP for any kind of real-time wire-line and wireless Internet services (conversational and streaming)

[6] IBBBP refers to Hierarchical B frame encoding

- File formats, e.g. ISO “MP4” for storage
- MPEG-2 systems for broadcasting services, etc.

One key concept of the NAL is parameter sets. A parameter set contains information that is expected to rarely change over time. There are two types of parameter sets: Sequence Parameter Sets (SPSs) apply to a series of consecutive coded video pictures; and Picture Parameter Sets (PPS), which apply to the decoding of one or more individual pictures. Inserting SPS and PPS headers before every IDR picture or I picture at a recovery point is essential to allow a decoder to begin playing from these points when they first open up a video that is being streamed. Without the information carried in these headers a decoder cannot begin decoding the video sequence.

6.2 Random Access Points

The H.264/AVC standard specifies that each sequence must have at least one IDR picture. The IDR picture (Instantaneous Decoder Refresh picture) provides a random access point from which a decoder can begin decoding.

When H.264/AVC video is streamed over a network, IDR pictures need to appear in the stream regularly to allow a decoder that connects to the stream to begin decoding. The time between IDR pictures determines the maximum time between connection to the stream and a decoder being able to begin decoding.

The international DVB standard [5] recommends that the random access points occur no less frequently than once every 5 seconds, and that for broadcast applications they occur at least once every 500 milliseconds. In practice, the application will determine the optimal GOP size.

Some encoder manufacturers prefer to use recovery point SEIs (Supplemental Enhancement Information) instead of IDRs to allow a stream to be decoded after a random jump to an arbitrary spot in stream. These recovery points should begin with an I -picture. All pictures in-display-order after the recovery point should not reference pictures in-decoding -order prior to the recovery point. Pictures in-display-order prior to the recovery point may reference pictures in the previous GOP that are in-display-order after their own GOP’s recovery point. Presence of an I picture and a recovery point SEI will provide similar random access points as those provided by IDRs.

6.3 PTS values for every video frame

Decoding Time Stamps (DTS) and Presentation Time Stamps (PTS) carried in the packetized elementary stream must be accurate as defined in the MPEG systems standard [3]. The PTS values provide information necessary to synchronize the decoding and display of the packetized elementary streams (video, metadata, audio, etc.) found in the transport stream.

The standard specifies that successive Presentation Time Stamp (PTS) values must not differ by more than 0.7 seconds for each Packetized Elementary Stream. However, it is recommended that there be a one-to-one association of a PTS to each video frame. This will ensure tight synchronization between video and synchronous metadata. PTS values can also provide the basis

for file navigation, so access to a particular frame is possible if specified to frame increments.

6.4 PTS values must occur no more than 10 seconds ahead of the corresponding PCR

The MPEG systems standard [3] specifies that all bytes for a given access unit must occur in the stream prior to the PCR time at which they must be decoded and displayed. Another way of saying this is that each PTS value must occur sufficiently ahead of the PCR so that all the bytes of the access unit will be available when the PCR clock reaches the PTS time. This ensures that all bytes of a video frame are available for decoding before the video frame is to be displayed.

Conversely, each video picture should not occur too far ahead of its display time, as this will contribute to increased latency. Even when latency is not an issue, it is required that each video picture not occur more than 10 seconds ahead of its display time. This will help to prevent buffer overflow in decoders which may have limited memory resources.

When the PTS value is too far head of the PCR or gets behind the PCR, the video playback may stutter, freeze, or result in poor video/audio synchronization, and can result in the inability to re-stream the content.

6.5 Fixed frame rate

The H.264/AVC standard allows the video frame rate to be fixed or variable. If fixed, the frame rate is constant for the duration of the sequence and this provides the greatest compatibility with existing systems.

If the video stream is being used for different types of tasks, such as both tracking and surveillance with critical imaging, allowing the frame rate to change in the stream from one rate to another provides advantages.

When the frame rate is changed, the change must be signaled with a new SPS, for example, switching from 30 frames-per-second to 15 frames-per-second.

For low latency modes, variable frame rate may be used at the expense of compatibility with current exploitation systems. Typical video encoding requires tradeoffs in picture quality, picture spatial resolution, temporal frame rate, and bit rate. Typically bit rate and spatial resolution are fixed. If frame rate is also fixed, the picture quality will vary depending on scene complexity and motion. In an attempt to maintain a constant picture quality, the frame rate can be allowed to vary as scene complexity and motion change. The latter mode has significant value in practical applications where picture quality is important; for example, in target recognition.

6.6 Fixed frame size

The H.264/AVC standard allows the video frame size (also known as video resolution) to change with a new SPS; for example, switching from 704x480 pixels to 352x240 pixels. Such a change may not be supported by some decoders.

6.7 Frame pattern

There are five main frame patterns used in H.264/AVC:

- I-frame only
- I-P which consists of just I and P frames
- I-B-P which includes I, B and P frames frequently using 2 B frames between each I or P in display order
- Hierarchical B, which may include more B frames between the I and P frames and additionally uses some B frames as references for other B frames
- Infinite GOP or P-frame only

I-frame only is the least efficient encoding pattern, but at high bit rates can provide the greatest encoding fidelity. It also allows fast random access and provides good error resilience since the errors do not propagate frame-to-frame.

I-P pattern is commonly used for low latency because it provides reasonable compression without introducing any frame re-ordering since the encoding order is the same as the display order.

I-B-P pattern is widely used as it provides good compression, but incurs a frame-reordering delay during encoding and decoding.

Hierarchical B frames coding structure provides very good compression and is gaining wider support.

Infinite GOP or P-frame only structures use a technique known as rolling macroblock refresh to remove the need for I frames. In effect the I-frame is distributed over a number of P frames. Infinite GOP or P-frame only is not widely supported but can provide the lowest latency form of encoding (along with I-frame only) due to the ability to use relatively fixed frame sizes and a constrained buffer model.

6.8 Picture Order Count (POC) equal to field number

Setting the Picture Order Count (POC) to equal the field number improves the ability for a file to be frame-accurately navigated and edited and is **recommended**.

6.9 No long-term reference pictures

While the use of long-term reference pictures can improve coding efficiency, they are not suitable for real-time streaming and largely prevent files from being randomly decoded and edited.

6.10 Video Profile

H.264/AVC supports a number of different profiles which are described in more detail in Appendix 8.1. Of these, Baseline, Main and High profile are the most commonly supported to date.

- Baseline, Main and High profiles are **recommended** for low latency applications.
- Main and Simple Baseline profiles are **recommended** for exploitation and dissemination.
- Baseline and Main profiles are **recommended** for low bandwidth applications.

6.11 Picture Not Coded

This is done by setting the coding mode for each macroblock in a P picture to **P_skip**. There are flags, **mb_skip_run** (CAVLC entropy coding) and **mb_skip_flag** (CABAC entropy coding) to indicate skipped macroblocks.

Definition of a **skipped macroblock**: A *macroblock* for which no data is coded other than an indication that the *macroblock* is to be decoded as "skipped". This indication may be common to several *macroblocks*.

If all macroblocks in a picture are skipped and, for example, CAVLC entropy coding mode is used, **mb_skip_run** can be set to the number of macroblocks in the entire picture indicating that all macroblocks in the picture are skipped. A frame coded in such a way will be coded using between 200 and 500 bits, depending on the format.

6.12 Additional syntax elements

This section contains several suggestions for additional information to be carried in the compressed data stream. This information is useful for decoders, editors and exploitation systems that may need to process the compressed video stream.

6.12.1 VUI parameters

The H.264/AVC standard includes optional Video Usability Information parameters. As its name implies, these parameters contain information that makes the video useable beyond simple decoding. While it may be possible to deduce some of the information in these parameters from the video and transport layer through numerical and statistical analysis, the VUI parameters provide an explicit and more reliable mechanism.

One of the more useful entries in the VUI parameters is the specification of the frame rate. The frame rate can be specified using the **time_scale** and **num_units_in_ticks** variables and setting **fixed_frame_rate_flag** to 1.

MaxFPS = Ceil(**time_scale** / (2 x **num_units_in_ticks**)) [see[1] EQ D-2]

Example: if time_scale = 60 000 and num_units_in_ticks = 1001, then MaxFPS = 29.97 Hz.

The **num_reorder_frames** parameter provides crucial information to reduce DPB output delay. The difference with or without this parameter can be significant for low latency applications and frame accurate seeking/editing applications.

The **max_dec_frame_buffering** parameter specifies the required size of the DPB in units of frame buffers. Many encoders actually use a **max_dec_frame_buffering** less than **MaxDpbSize**; by providing this information, decoders can reduce Decoder Picture Buffer (DPB) size and latency (when **num_reorder_frames** is not present, its value is inferred to be equal to **max_dec_frame_buffering**).

Note: one frame's worth of DPB saving is much more significant than one frame's worth of CPB (Coded Picture Buffer) saving since the former is in raw, uncompressed format.

6.12.2 Use of HRD parameters

The HRD parameters define the buffer sizes and bit rates of the operation of the Hypothetical Reference Decoder (HRD) for the bit stream. These parameters are critical if the video stream is to be multiplexed with other elementary streams.

The **cpb_size_value_minus1** parameter is a very important for editing or multiplexing procedures, because it directly affects the PCR/PTS gap of resulting footage. In practice, most encoders choose CPB buffer size to be much smaller than its maximum value restricted by level ([1] Annex A), in order to make their output playable on the widest range of hardware players.

6.12.3 Buffering period SEI message

The Buffering period SEI message occurs before each IDR and provides the initial buffering requirement, at random access points. The purpose of the initial buffering is to keep the encoder and decoder buffer levels complementary. The decoder ignores subsequent buffering period SEI messages once it begins decoding.

6.12.4 Other video parameters

The **gaps_in_frame_num_value_allowed_flag** should be set to 0 (gaps not allowed).

The **aspect_ratio_info_present_flag** shall be set to 1 and the **aspect_ratio_idc** shall be correctly coded for the sample aspect ratio of the video content.

6.13 Picture timing SEI message

Reference MISB STANDARD 0604 [10]

7 Transport Stream Multiplexing Recommended Practices

While there are a number of transport formats that allow H.264/AVC video to be carried together with audio, metadata and other elementary streams, there are two which are most commonly used and will provide the greatest compatibility with existing systems and software: the MPEG-2 Transport stream and the Real Time Protocol (RTP).

The H.264/AVC specification [1] distinguishes conceptually between a Video Coding Layer (VCL), and a Network Abstraction Layer (NAL). The VCL contains the video features of the codec (transform coefficients, quantization and motion information, loop filter parameters, etc.). The NAL layer formats the VCL data into Network Abstraction Layer units (NAL units) suitable for transport across the applied network or storage medium. A NAL unit consists of a one-byte header and the payload; the header indicates the type of the NAL unit and other information, such as the (potential) presence of bit errors or syntax violations in the NAL unit payload, and information regarding the relative importance of the NAL unit for the decoding process.

The NAL units can be placed into a bit stream with the addition of start code bytes. This mode is used for putting H.264/AVC into an MPEG-2 Transport Stream. Similarly, the NAL units can be placed into packets following the RTP or other packet based protocols.

7.1 Carriage over MPEG2 Transport Stream

MPEG-2 Transport Stream can be used for H.264/AVC video transmission over serial communications links such as radio, satellite, telecom, and broadcast systems. Carriage of H.264/AVC video in MPEG-2 transport streams is defined in [4].

MPEG-2 Transport Stream can also be used for H.264/AVC video transmission over TCP/IP networks when the raw transport stream packets are placed into UDP packets.

The following transport stream guidelines are a subset of the most widely mandated requirements of the broadcast industry (e.g. from such standards as the CableLabs Specification). Conformance with these guidelines will maximize the likelihood that a transport stream can be processed by commercial off-the-shelf (COTS) hardware and software, including storage, transmission, multiplexing, and decoding.

- Encapsulate every video AU into its own PES packet.
- Every PES header shall contain the PTS (and DTS, if required) of the video AU contained in the PES packet.
- PCRs must occur with a separation of less than 100 milliseconds.
- Encoder PCR accuracy at 27 MHz must be +/- 5ppm [5].
- A Program Association Table (PAT) for the program must occur in the transport stream before any Program Map Table (PMT) for the program.
- Both PAT and PMT must be inserted in the transport stream greater than 4 times per

second (8 times per second **recommended**) throughout the program to allow rapid program acquisition.

- Transport packet at the start of an H.264/AVC RAP (Random Access Point: IDR picture or I picture with recovery point SEI, SPS and PPS from which video decoding can begin successfully) should have **random_access_indicator** set to 1.
- In support of future trick modes, the **elementary_stream_priority_indicator** bit shall be set whenever an access unit containing an I-picture is present in H264/AVC video streams.
- It is **recommended** that any error detecting devices in the transmission path set the **transport_error_indicator** bit in the transport packet when uncorrectable errors are detected. At the decoder, if this flag is set then suitable concealment or error recovery measures can be effected.

7.2 Carriage over RTP

Real Time Protocol (RTP) can be used for H.264/AVC video transmission over IP networks. RFC 3984 [7] specifies how to carry H.264/AVC NAL units in RTP packets.

Use of RTP requires the definition of payload formats that are specific for each content format, and so the system layer specifies which RTP payload formats to use for transport of advanced audio and video, as well as applicable constraints. For transport over IP, the H.264/AVC data is packetized in RTP packets using RFC 3984, and with restrictions as specified in RP 0804 [9].

8 Appendix A

This appendix includes additional discussions related to different encoding options.

8.1 H.264/AVC Profiles and Levels

The H.264/AVC Standard is designed to be generic in that it serves a wide range of applications, bit rates, resolutions, qualities, and services. As a result the standard consists of a common syntax which carries a large number of optional “tools” or syntax elements. These tools provide various video compression and processing functions. Not every application needs, or can benefit from every tool. Nor is it practical for a decoder to implement and be able to support every tool. For this reason, limited subsets of the tools are grouped by means of "profiles" and "levels".

A "profile" is a subset of syntax elements, loosely grouped for an expected application. A “level” is a specified set of constraints imposed on values of the syntax elements in the bitstream.

A decoder that declares support for a specific profile and level must support all of the tools and constraints defined therein. An encoder that declares support for a specific profile and level may not utilize any tools from outside that profile, nor can it exceed the constraints on the tools that are allowed. However, an encoder need not utilize all of the tools nor must it operate up to the limit of the constraints. An encoder is compliant as long as it generates a legal bitstream syntax within the profile and level that it is operating.

The following subsections discuss some of the features of the common profiles and levels.

8.1.1 Baseline Profile

The Baseline Profile contains the following restricted set of coding features.

- I and P Slices: Intra coding of macroblocks through the use of I slices; P slices add the option of Inter coding using one temporal prediction signal. Baseline profile does not support the use of B Slices.
- 4x4 Transform: The prediction residual is transformed and quantized using 4x4 blocks.
- CAVLC: The symbols of the coder (e.g. quantized transform coefficients, intra predictors, motion vectors) are entropy-coded using a variable length code.

In addition, the Baseline profile includes coding features which support error resilience. This is useful for video transmission in error prone environments. Care should be taken in specifying the use of these tools, since they may not be widely supported.

- Flexible Macroblock Ordering (FMO): This feature allows arbitrary sampling of the macroblocks within a slice.
- Arbitrary Slice Ordering (ASO): This feature allows arbitrary order of slices within a picture.

- Redundant Slices: This feature allows transmission of redundant slices that approximate the primary slice.

8.1.2 Constrained Baseline Profile

The Constrained Baseline Profile contains features that in are common between the Baseline and Main profiles.

8.1.3 Main Profile

Main Profile contains all features of Baseline Profile, except for FMO, ASO, and Redundant Slices, plus the following additional features:

- B Slices: Enhanced Inter coding using up to two temporal prediction signals that are superimposed for the predicted block.
- Weighted Prediction: Allowing the temporal prediction signal in P and B slices to be weighted by a factor.
- CABAC: Alternative entropy coding to CAVLC providing higher coding efficiency at higher complexity, which is based on context-adaptive binary arithmetic coding.
- Field Encoding: Field encoding provides for more efficient compression of interlaced video particularly in the presence of motion. While the goal is to move toward progressive sensors, there are and will continue to be interlaced sensors.

8.1.4 High Profile

High Profile contains all features of Main Profile and the following additional ones:

- 8x8 Transform: In addition to the 4x4 Transform, the encoder can choose to code the prediction residual using an 8x8 Transform.
- Quantization Matrices: The encoder can choose to apply different weights to the transform coefficients. This allows better alignment of quantization with human perception.
- Monochrome/single channel/4:0:0
- Larger allowable level prefix in CAVLC
- Separate C_b and C_r chroma QP control

8.1.5 High 10 Profile

High 10 Profile contains all of the features of High Profile with the addition of support for 9 and 10-bit pixel depths. Care should be taken in specifying the use of these tools, since they may not be widely supported.

8.1.6 High 4:2:2 Profile

High 4:2:2 Profile contains all of the features of High 10 Profile with the addition of support for 4:2:2 chroma sampling. Care should be taken in specifying the use of these tools, since they may not be widely supported.

8.1.7 High 4:4:4 Profile

High 4:4:4 Profile contains all of the features of High 4:2:2 Profile with the following additions:

- Support for up to 14-bit pixel depths.
- Support for 4:4:4 chroma sampling
- Residual color transform
- Predictive lossless coding

Care should be taken in specifying the use of these tools, since they may not be widely supported.

8.2 Levels

Levels typically constrain picture size, frame rate, and bitrate. Table A1, while not a complete definition of the levels, summarizes the key features for baseline, extended and main profile. Bit rates are for video coding layer.

Table A1 – H.264 Baseline, Extended, and Main profile levels

Level Number	Typical Picture Size	Typical Frame Rate Maximum	Compressed Bit Rate Maximum
1	128x96p 176x144p	30 15	64 kbps
1b	176x144p	15	128 kbps
1.1	176x144p 352x288p	30 7.5	192 kbps
1.2	352x288p	15	384 kbps
1.3	352x288p	30	768 kbps
2	352x288p	30	2 Mbps
2.1	352x480p/i 352x576p/i	30 25	4 Mbps
2.2	720x480p/i 720x576p/i	15 12.5	4 Mbps
3	720x480p/i 720x576p/i	30 25	10 Mbps
3.1	1280x720p	30	14 Mbps
3.2	1280x720p	60	20 Mbps
4	1280x720p 1920x1080p	60 30	20 Mbps
4.1	1280x720p	60	50 Mbps

	1920x1080p	30	
4.2	1920x1080p	60	50 Mbps
5	2048x1024p	72	135 Mbps
5.1	2048x1024p 4096x2048p	120 30	240 Mbps

Table Notes: suffix "p" denotes progressive frame (i.e. 480p30 means 30 frames where each frame contains 480 lines imaged at a 30 Hz rate). Suffix "i" denotes interlaced frame (i.e. 480i30 means 30 complete frames per second constructed from pairs of fields taken at 60 fields per second). A 30 frames-per-second sequence constructed from pairs of 60 fields per second interlaced fields is *not* equivalent to a 30 frames per second progressive sequence (except in the simple case of imagery without any motion). The latter is imaged by the sensor as one complete frame, while the former is imaged as two separate fields displaced temporally in time by the field rate.

8.3 Group of Pictures (GOP) Length

A Group of Pictures (GOP) consists of all the pictures from one I picture (represented by a random access point) to the picture just prior to the next I picture. The GOP Length will represent a trade-off between the needs of random access and error recovery versus bandwidth.

Smaller GOP Lengths are desirable for fine grain random access. This is because random access to a picture which occurs between random access points must start at either the prior or the next random access point. Similarly, in streaming applications the decoder must discard pictures until it receives a random access point to begin decoding the stream.

Smaller GOP Lengths are also desirable in error prone environments, since a video artifact caused by an error in the transmitted data stream will often persist in the decoded video until the next I picture.

Since I pictures are roughly 5-10 times larger than P pictures, reducing the number of I pictures will improve the efficiency of the compression and allow better quality video (assuming no transmission errors) at the same bandwidth. When video is transmitted over constant bitrate communications channels, as is the case in many radio links, GOP Length will have a direct impact on video quality. Using small GOP Lengths will increase the number of I pictures, reducing video quality, while larger GOP Lengths will decrease the number of I pictures, increasing the video quality.

For low latency operator-in-the-loop applications, an infinite GOP Length sequence can help reduce the latency at the expense of interoperability.

In infinite GOP mode a single I picture is transmitted at the start of the sequence, and then all other pictures are coded as P pictures. In this mode, there are no random access points. The benefit of this mode is twofold. First, since all the coded pictures are roughly the same size (excluding the first I picture), there is a significant reduction in the buffering that is required in the encoder and the decoder during constant bitrate operation. This reduction in buffering translates into a reduction in latency which is critical in many real time applications. Second, since no I pictures are transmitted, only more efficient P pictures are used. In addition to the loss

of random access, this mode also is very susceptible to transmission errors which could persist forever if unchecked. To correct this, intra refresh is used. In intra refresh, a small number of blocks are intra coded in each picture. The intra coded blocks do not use temporal prediction just like the blocks in I pictures, correcting any transmission errors that may have occurred. By selecting the number of intra coded blocks per picture, you can select the number of pictures over which the entire picture will be refreshed. The SPS and PPS can still be sent periodically to aid decoders which receive the stream in mid sequence.

8.4 Slices

Slices provide resynchronization points within a picture. Because modern compression algorithms make heavy use of variable length coding, a single bit error can have a significant impact on video artifacts. In some cases, this can corrupt the picture from the point of the error to the end of the frame. Slices provide resynchronization points within a frame so that an error will only corrupt the picture to the end of the current slice.

While slices add additional overhead, they can be useful in high error environments. Normal operation uses a single slice which contains the whole picture. User control of slices is usually controlled by specifying a slice size in bytes, or a slice size in macroblocks. Specifying the slice size in bytes allows the slice to be aligned with transport structures such as an RTP packet or a Forward Error Correction frame. Specifying the slice size in macroblocks aligns the slices to the picture structure.

8.5 Quantizer

The Quantizer controls the bitrate and video quality. The H.264/AVC encoder will typically vary the quantizer used throughout the picture in order to maintain the highest video quality while not exceeding the target data rate. Larger quantizer values typically will cause more information to be discarded or reproduced less accurately. Lower quantizer values attempt to preserve information and reproduce it more accurately. Normally, the perceptually less important high frequency information is discarded first. This results in blurriness or fuzziness in the image. Preservation of the high frequency information results in improved detail and sharpness of edges.

To maintain higher quality video at a given bitrate it is more efficient to reduce the frame rate than to just let the quantizer values increase, assuming lower frame rate is acceptable.

8.6 Frame Rate

Some encoders allow the user to encode the video at a lower frame rate than the input source. This is useful when picture quality is more important than motion reproduction or latency. By reducing the frame rate, the user is allowing the algorithm to allocate more bits to each picture, resulting in higher quality pictures.

8.7 B Frames

While P-Frames are predicted from the previous picture, B-Frames are predicted from both previous pictures and future pictures. B-Frames provide the most efficient way to encode a frame however this efficiency decreases with the frame distance of the B frame from its reference frame. In H.264/AVC, unlike in MPEG-2, B frames can also be used as reference frames for other B frames. An encoding structure called Hierarchical B's is becoming increasingly popular as it increases the ratio of B frames to the less efficient I and P frames.

B frames however are usually not suitable for low latency encoding because their reference frames need to be transmitted before the current B frame itself can be coded and the subsequent re-ordering of frames in the decoder also increases latency.

8.8 Field Encoding

The use of interlaced-scanned sensors is rejected by the MISP because interlaced scanning will introduce temporal artifacts into the imagery. However, should an encoder need to encode video from a legacy sensor, field encoding may produce better quality. Normally, the entire frame is treated as a single picture and divided up into macroblocks for processing by the compression algorithm. This is fine if the video is from a progressive camera which captures a whole frame at one time, or if the video scene has very little horizontal motion. But the video will exhibit interlace artifacts if the camera uses an interlaced sensor which captures fields at twice the frame rate (one field of odd lines followed by one field of even lines) and if there is horizontal motion. The interlace artifacts are the result of a moving object being in one position when the first field is captured and in a new position when the second field is captured. When you look at the image as a whole frame, the object edges appear to be torn since every other line (from opposite fields) represents a different point in time. The result is a lot of very high frequency information in the image which is very inefficient to compress. At lower resolutions where only a single field is used, this is not an issue.

Field encoding separates the frame into the two fields and compresses them separately. This eliminates the high frequency information due to the interlace artifacts and significantly improves the compression efficiency.

There are two methods of doing field encoding with H.264/AVC. Macroblock Adaptive Frame/Field (MBAFF) makes the frame versus field encoding decision on a block by block basis throughout the picture, applying the most efficient method to each block as needed. Picture Adaptive Frame/Field (PAFF) makes the frame versus field encoding decision on a frame by frame basis coding the entire frame using the same method. Field encoding is not available in Baseline Profile.

8.9 Entropy Coding

Entropy coding is the last step in the compression process. This is a coding process which converts the compression coefficients and other data into an efficient format for transmission. H.264/AVC provides two types of entropy coding. One method is Context-adaptive variable-

length coding (CAVLC). The other method is Context-Based Adaptive Binary Arithmetic Coding (CABAC), and is only available in Main and High Profiles. CABAC is noticeably more efficient than CAVLC, with an average bitrate reduction of 10-15%; however, it is more computationally intensive. CABAC may be especially useful at very low data rates (< 1Mbps). CABAC is not available in Baseline Profile.

8.10 Rate Control

Rate Control defines the data rate characteristics of the generated compressed data. There are typically three rate control modes. First is Constant Bit Rate (CBR). In this mode, the encoder will control the compression algorithm so that an approximately constant number of bits are output over any short time period in order to not exceed the selected data rate. Fill data is added, if necessary, to avoid under running the selected data rate. This mode would typically be used when transmitting video over a fixed data rate communications link.

The second mode is Variable Bit Rate (VBR). In this mode, the encoder will control the compression algorithm in order to maintain the data rate within some selected minimum and maximum values. The algorithm is able to allocate bits more flexibly, using fewer bits in less demanding scenes and more bits in difficult-to-encode scenes, resulting in higher video quality than for CBR mode at the same average bitrate. The disadvantages are that it may take more time to encode, as the process is more complex, and that some hardware might not be compatible with VBR data. This mode is applicable to variable data rate communications links such as an Ethernet network.

The third mode is Constant Quality (CQ), sometimes called unconstrained VBR. In this mode, the encoder controls the compression algorithm in order to maintain the same quality for every frame. Quality is usually selected with the Quantizer parameter. This mode will attempt to produce the same quality for all frames and has no consideration of data rate generated. This mode is applicable to variable data rate communications links such as an Ethernet network.

In addition to the above rate control modes, there are also some encoders which are called Multi-Pass encoders. These encoders are typically not applicable to real time video streams since they must process the same sequence of frames several times. The benefit is that they can have very high compression efficiencies because they can analyze the video during early passes in order to apply the optimum compression techniques in later passes. The penalty for this efficiency is a significant increase in latency. These encoders, however, are very useful for offline compression.

9 Appendix B

Below is presented a theoretical model of expected sources of latency in a video encoder and decoder. This should aid in understanding how frame rate impacts latency, and the greatest contributors to latency in the encode/decode signal path. (Provided by Dr. Isnardi, Sarnoff Corp)